SORT 38 (2) July-December 2014, 305-324

Global hypothesis test to compare the likelihood ratios of multiple binary diagnostic tests with ignorable missing data

Ana Eugenia Marín-Jiménez[1] and José Antonio Roldán-Nofuentes[1]

Abstract

In this article, a global hypothesis test is studied to simultaneously compare the likelihood ratios of multiple binary diagnostic tests when in the presence of partial disease verification the missing data mechanism is ignorable. The hypothesis test is based on the chi-squared distribution. Simulation experiments were carried out to study the type I error and the power of the global hypothesis test when comparing the likelihood ratios of two and three diagnostic tests respectively. The results obtained were applied to the diagnosis of coronary stenosis.

MSC: 62P10 (Applications to biology and medical science), 6207 (Data analysis)

Keywords: Global hypothesis test, partial verification, positive and negative likelihood ratio.

1. Introduction

The fundamental parameters to assess the accuracy of a binary diagnostic test are the sensitivity and the specificity. Sensitivity (Se) is the probability of the diagnostic test being positive when the individual has the disease, and specificity (Sp) is the probability of the diagnostic test being negative when the individual does not have the disease. Sensitivity and specificity depend on the intrinsic ability of the diagnostic test to distinguish between individuals with and without the disease. Other parameters to assess the accuracy of a binary diagnostic test are the likelihood ratios (LRs). When the result of the diagnostic test is positive, the likelihood ratio, called positive likelihood ratio (LR^+), is the ratio between the probability of a positive test result in individuals with the disease (Se) and the probability of a positive result in individuals without the disease $(1-Sp)$. When the result of the diagnostic test is negative, the likelihood ratio,

[1] Biostatistics, Department of Statistics, University of Granada, Spain, email: anamarin@ugr.es, jaroldan@ugr.es
Received: March 2014
Accepted: June 2014

called negative likelihood ratio (LR^-), is the ratio between the probability of a negative test result in individuals with the disease $(1-Se)$ and the probability of a negative test result in individuals without the disease (Sp). The *LRs* only depend on the sensitivity and specificity of the diagnostic test, and their values vary between zero and infinite. When the diagnostic test and the gold standard are independent then $LR^+ = LR^- = 1$, and if the diagnostic test correctly classifies all of the individuals (with or without the disease) then $LR^+ = \infty$ and $LR^- = 0$. A value of $LR^+ > 1$ indicates that a positive test result is more probable for an individual with the disease than for an individual without the disease, and a value of $LR^- < 1$ indicates that a negative test result is more probable for an individual who does not have the disease than for one who has the disease. The *LRs* quantify the increase in knowledge of the disease presence through the application of the diagnostic test. Let T be the random variable that models the result of the diagnostic test ($T = 1$ when the result is positive and $T = 0$ when the result is negative), let D be the random variable that models the result of the gold standard ($D = 1$ when the individual is diseased and $D = 0$ when this is not the case), and $p = P(D = 1)$ the disease prevalence in the population which is subject to the study. The ratio between the probability that an individual has the disease and the probability that an individual does not have the disease before applying the diagnostic test is

$$\text{Odds pre-test} = \frac{p}{1-p}.$$

After applying the diagnostic test the ratio is

$$\text{Odds post-test}(T) = \frac{P(D=1|T)}{P(D=0|T)}.$$

The *LRs* relate the two previous odds, i.e.

$$\text{Odds post-test}(T=1) = LR^+ \times \text{Odds pre-test}$$

and

$$\text{Odds post-test}(T=0) = LR^- \times \text{Odds pre-test}$$

Furthermore, the comparison of the *LRs* of diagnostic tests has been the subject of different studies. In designs with independent samples, Luts et al (2011) studied a hypothesis test to compare the *LRs* of two or more binary diagnostic tests studying the effect of sample sizes on the asymptotic behaviour of the proposed test. The hypothesis test proposed by these authors allows us to simultaneously compare the *LRs* of the diagnostic tests subject to this type of sample design and is based on the chi-squared distribution. For paired designs, Leisenring and Pepe (1998) proposed a *GEE* model

to independently compare the positive *LRs* and the negative *LRs* of two diagnostic tests; and Roldán Nofuentes and Luna del Castillo (2007) proposed a hypothesis test to independently and jointly compare the positive *LRs* and the negative *LRs* of two diagnostic tests through a likelihood-based approach. Nevertheless, in clinical practice the gold standard is frequently not applied to all of the individuals in a sample, leading to the problem known as partial disease verification (Begg and Greenes, 1983; Zhou, 1993). In this situation, the disease status (whether the disease is present or absent) is unknown for a subset of individuals in the sample, and therefore if the previous parameters are estimated only considering those individuals whose disease status are known, the estimators are affected by what is known as verification bias. The same problem occurs when, in the presence of partial disease verification, the parameters of two (or more) binary diagnostic tests are compared in relation to the same gold standard. When in the presence of partial verification the missing data mechanism is MAR, Roldán Nofuentes and Luna del Castillo (2005) studied a hypothesis test to independently compare the *LRs* of two binary diagnostic tests. In this article, we extend the results of these authors and we study a hypothesis test to simultaneously compare the *LRs* of two or more binary diagnostic tests. In Section 2, we propose a global hypothesis test based on the chi-squared distribution to simultaneously compare the *LRs* of multiple binary diagnostic tests when, in the presence of partial disease verification, the missing data is ignorable. In Section 3, we carry out simulation experiments to study the type I error and the power of the global hypothesis test when simultaneously comparing the *LRs* of two and of three binary diagnostic tests. In Section 4, the global test is applied to an example and in Section 5 we discuss the results obtained.

2. Global hypothesis test

Let us consider J binary diagnostic tests ($J \geq 2$) that are applied independently to all of the individuals in a random sample sized n, and a gold standard that is only applied to a subset of the n individuals in the sample. Let T_j ($j = 1,\ldots,J$), V and D be the random variables defined as: T_j models the result of the jth binary test ($T_j = 1$ when the test result is positive and $T_j = 0$ when it is negative); V models the verification process ($V = 1$ when the individual is verified with the gold standard and $V = 0$ when the individual is not verified with the gold standard); and D models the result of the gold standard ($D = 1$ when the individual has the disease and $D = 0$ when the individual does not have the disease). Let $s_{i_1,\ldots,i_J}$ be the number of individuals verified in which $T_1 = i_1$, $T_2 = i_2,\ldots,T_J = i_J$ and $D = 1$; $r_{i_1,\ldots,i_J}$ the number of individuals verified in which $T_1 = i_1$, $T_2 = i_2,\ldots,T_J = i_J$ and $D = 0$; and $u_{i_1,\ldots,i_J}$ the number of individuals not verified in which $T_1 = i_1$, $T_2 = i_2,\ldots,T_J = i_J$, with $i_j = 0, 1$ and $j = 1,\ldots,J$. Let $n_{i_1,\ldots,i_J} = s_{i_1,\ldots,i_J} + r_{i_1,\ldots,i_J} + u_{i_1,\ldots,i_J}$ and $n = \sum_{i_1,\ldots,i_J=0}^{1} n_{i_1,\ldots,i_J}$. Let the probabilities be

$$\phi_{i_1,\ldots,i_J} = P(V=1, D=1, T_1=i_1, \ldots, T_J=i_J)$$
$$\varphi_{i_1,\ldots,i_J} = P(V=1, D=0, T_1=i_1, \ldots, T_J=i_J)$$

and

$$\gamma_{i_1,\ldots,i_J} = P(V=0, T_1=i_1, \ldots, T_J=i_J)$$

with $i_j = 0, 1$, and it is verified that

$$\sum_{i_1,\ldots,i_J=0}^{1} \phi_{i_1,\ldots,i_J} + \sum_{i_1,\ldots,i_J=0}^{1} \varphi_{i_1,\ldots,i_J} + \sum_{i_1,\ldots,i_J=0}^{1} \gamma_{i_1,\ldots,i_J} = 1.$$

Let $\boldsymbol{\omega} = (\phi_{1,\ldots,1}, \ldots, \phi_{0,\ldots,0}, \varphi_{1,\ldots,1}, \ldots, \varphi_{0,\ldots,0}, \gamma_{1,\ldots,1}, \ldots, \gamma_{0,\ldots,0})^{\mathsf{T}}$ be a vector of size $3 \cdot 2^J$ whose components are the previous probabilities. As the disease status of all the individuals in the sample is not verified with the gold standard, the verification probabilities are defined as

$$\lambda_{k,i_1,\ldots,i_J} = P(V=1|D=k, T_1=i_1, T_2=i_2, \ldots, T_J=i_J).$$

Therefore, $\lambda_{k,i_1,\ldots,i_J}$ is the probability of selecting an individual to verify the disease status in which D=k, $T_1=i_1$, $T_2=i_2, \ldots$ and $T_J=i_J$, with $k, i_j = 0, 1$, $j=1,\ldots,J$. If the verification process only depends on the results of the J binary tests and does not depend on the disease status, that is to say when $\lambda_{k,i_1,\ldots,i_J} = \lambda_{i_1,\ldots,i_J} = P(V=1|T_1=i_1, T_2=i_2, \ldots, T_J=i_J)$, this is equivalent to assuming that the verification process is missing at random (MAR) (Rubin, 1976). Assuming that the verification process is MAR and that the parameters of the data model and the parameters of the missingness mechanism are different, the missing data mechanism is called to be ignorable (Schafer, 1997) and all of the parameters of the model can be estimated applying the method of maximum likelihood. Under this assumption, the *LRs* of the *j*th diagnostic test are written as

$$LR_j^+ = \frac{(1-p)\left(\sum\limits_{\substack{i_1,\ldots,i_J=0 \\ i_j=1}}^{1} \frac{\phi_{i_1,\ldots,i_J}\eta_{i_1,\ldots,i_J}}{\phi_{i_1,\ldots,i_J}+\varphi_{i_1,\ldots,i_J}}\right)}{p\left((1-p) - \sum\limits_{\substack{i_1,\ldots,i_J=0 \\ i_j=0}}^{1} \frac{\varphi_{i_1,\ldots,i_J}\eta_{i_1,\ldots,i_J}}{\phi_{i_1,\ldots,i_J}+\varphi_{i_1,\ldots,i_J}}\right)}$$

and

$$LR_j^- = \frac{(1-p)\left(p - \sum\limits_{\substack{i_1,\ldots,i_J=0 \\ i_j=1}}^{1} \frac{\phi_{i_1,\ldots,i_J}\eta_{i_1,\ldots,i_J}}{\phi_{i_1,\ldots,i_J}+\varphi_{i_1,\ldots,i_J}}\right)}{p\left(\sum\limits_{\substack{i_1,\ldots,i_J=0 \\ i_j=0}}^{1} \frac{\varphi_{i_1,\ldots,i_J}\eta_{i_1,\ldots,i_J}}{\phi_{i_1,\ldots,i_J}+\varphi_{i_1,\ldots,i_J}}\right)},$$

where $p = \sum\limits_{i_1,\ldots,i_J=0}^{1} \frac{\phi_{i_1,\ldots,i_J}\eta_{i_1,\ldots,i_J}}{\phi_{i_1,\ldots,i_J}+\varphi_{i_1,\ldots,i_J}}$ is the disease prevalence and $\eta_{i_1,\ldots,i_J} = \phi_{i_1,\ldots,i_J} + \varphi_{i_1,\ldots,i_J} + \gamma_{i_1,\ldots,i_J}$. The log-likelihood of the observed data is

$$l = \sum_{i_1,\ldots,i_J=0}^{1} s_{i_1,\ldots,i_J}\log(\phi_{i_1,\ldots,i_J}) + \sum_{i_1,\ldots,i_J=0}^{1} r_{i_1,\ldots,i_J}\log(\varphi_{i_1,\ldots,i_J}) + \sum_{i_1,\ldots,i_J=0}^{1} u_{i_1,\ldots,i_J}\log(\gamma_{i_1,\ldots,i_J})$$

Maximizing this function, the maximum likelihood estimators of the probabilities $\phi_{i_1,\ldots,i_J}$, $\varphi_{i_1,\ldots,i_J}$ and $\gamma_{i_1,\ldots,i_J}$ are

$$\hat{\phi}_{i_1,\ldots,i_J} = \frac{s_{i_1,\ldots,i_J}}{n}, \hat{\varphi}_{i_1,\ldots,i_J} = \frac{r_{i_1,\ldots,i_J}}{n} \text{ and } \hat{\gamma}_{i_1,\ldots,i_J} = \frac{u_{i_1,\ldots,i_J}}{n}$$

and, therefore, the maximum likelihood estimators of the *LRs* of the *j*th diagnostic test are

$$\widehat{LR}_j^+ = \frac{\left(\sum\limits_{i_1,\ldots i_J=0}^{1} \frac{r_{i_1,\ldots i_J}n_{i_1,\ldots i_J}}{s_{i_1,\ldots i_J}+r_{i_1,\ldots i_J}}\right)\left(\sum\limits_{\substack{i_1,\ldots i_J=0 \\ i_J=1}}^{1} \frac{s_{i_1,\ldots i_J}n_{i_1,\ldots i_J}}{s_{i_1,\ldots i_J}+r_{i_1,\ldots i_J}}\right)}{\left(\sum\limits_{i_1,\ldots i_J=0}^{1} \frac{s_{i_1,\ldots i_J}n_{i_1,\ldots i_J}}{s_{i_1,\ldots i_J}+r_{i_1,\ldots i_J}}\right)\left(\sum\limits_{i_1,\ldots i_J=0}^{1} \frac{r_{i_1,\ldots i_J}n_{i_1,\ldots i_J}}{s_{i_1,\ldots i_J}+r_{i_1,\ldots i_J}} - \sum\limits_{\substack{i_1,\ldots i_J=0 \\ i_J=0}}^{1} \frac{r_{i_1,\ldots i_J}n_{i_1,\ldots i_J}}{s_{i_1,\ldots i_J}+r_{i_1,\ldots i_J}}\right)}$$

and

$$\widehat{LR}_j^- = \frac{\left(\sum\limits_{i_1,\ldots i_J=0}^{1} \frac{r_{i_1,\ldots i_J}n_{i_1,\ldots i_J}}{s_{i_1,\ldots i_J}+r_{i_1,\ldots i_J}}\right)\left(\sum\limits_{i_1,\ldots i_J=0}^{1} \frac{s_{i_1,\ldots i_J}n_{i_1,\ldots i_J}}{s_{i_1,\ldots i_J}+r_{i_1,\ldots i_J}} - \sum\limits_{\substack{i_1,\ldots i_J=0 \\ i_J=1}}^{1} \frac{s_{i_1,\ldots i_J}n_{i_1,\ldots i_J}}{s_{i_1,\ldots i_J}+r_{i_1,\ldots i_J}}\right)}{\left(\sum\limits_{i_1,\ldots i_J=0}^{1} \frac{s_{i_1,\ldots i_J}n_{i_1,\ldots i_J}}{s_{i_1,\ldots i_J}+r_{i_1,\ldots i_J}}\right)\left(\sum\limits_{\substack{i_1,\ldots i_J=0 \\ i_J=0}}^{1} \frac{r_{i_1,\ldots i_J}n_{i_1,\ldots i_J}}{s_{i_1,\ldots i_J}+r_{i_1,\ldots i_J}}\right)}.$$

Let $\boldsymbol{\eta} = \left(LR_1^+, \ldots, LR_J^+, LR_1^-, \ldots, LR_J^-\right)^\mathsf{T}$ be a vector of size $2J$ whose components are the *LRs* of each diagnostic test. As the vector ω is the vector of probabilities of a multinomial distribution, the variance-covariance matrix of $\hat{\boldsymbol{\omega}}$ is $\sum_{\hat{\omega}} = \left\{\text{diag}\left(\boldsymbol{\omega}\right) - \boldsymbol{\omega}\boldsymbol{\omega}^\mathsf{T}\right\}/n$, and applying the delta method (Agresti, 2002) the asymptotic variance-covariance matrix of $\hat{\boldsymbol{\eta}}$ is

$$\sum_{\hat{\boldsymbol{\eta}}} = \left(\frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{\omega}}\right) \sum_{\hat{\boldsymbol{\omega}}} \left(\frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{\omega}}\right)^\mathsf{T} \tag{1}$$

The positive and negative *LRs* of each one of the J diagnostic tests depend on the same parameters (sensitivity and specificity of the *jth* diagnostic test) and, therefore, these parameters can be compared simultaneously. The global hypothesis test to compare simultaneously the *LRs* of the J diagnostic tests is

$$H_0: LR_1^+ = LR_2^+ = \cdots = LR_J^+ \text{ and } LR_1^- = LR_2^- = \cdots = LR_J^-$$
$$H_1: \text{at least one equality is not true.}$$

This hypothesis test is equivalent to the hypothesis test

$$H_0 : \boldsymbol{\psi}\,\boldsymbol{\eta} = 0 \text{ vs } H_1 : \boldsymbol{\psi}\,\boldsymbol{\eta} \neq 0 \tag{2}$$

where $\boldsymbol{\psi}$ is a full range matrix whose dimension is $2\left(J-1\right) \times 2J$, and whose elements are known values. For $J = 2$ the matrix $\boldsymbol{\psi}$ is

$$\boldsymbol{\psi} = \begin{pmatrix} 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{pmatrix}$$

and for $J \geq 3$ the matrix $\boldsymbol{\psi}$ is

$$\boldsymbol{\psi} = \begin{pmatrix} \boldsymbol{\psi}_1 & \boldsymbol{\psi}_0 \\ \boldsymbol{\psi}_0 & \boldsymbol{\psi}_1 \end{pmatrix}$$

where $\boldsymbol{\psi}_0$ is a matrix $(J-1) \times J$ with all of the elements equal to 0, and $\boldsymbol{\psi}_1$ is a matrix $(J-1) \times J$ where the elements (i, i) are equal to 1, the elements $(i, i+1)$ are equal to -1 for $i = 1, \ldots, J-1$, and the rest of the elements in this matrix are equal to 0. Applying the multivariate central limit theorem it is verified that

$$\sqrt{n}\left(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}\right) \xrightarrow[n\to\infty]{} N_{2J}\left(\mathbf{0}, \sum_{\boldsymbol{\eta}}\right).$$

Then, the statistic $Q^2 = \left(\boldsymbol{\psi}\,\hat{\boldsymbol{\eta}}\right)^{\mathsf{T}}\left(\boldsymbol{\psi}\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\eta}}}\boldsymbol{\psi}^{\mathsf{T}}\right)^{-1}\boldsymbol{\psi}\,\hat{\boldsymbol{\eta}}$ is distributed according to Hotelling's T-squared distribution with a dimension 2(J-1) and n degrees of freedom, where $2(J-1)$ is the dimension of the vector $\boldsymbol{\psi}\,\hat{\boldsymbol{\eta}}$. When n is large, the statistic Q^2 is distributed according to a central chi-squared distribution with 2(J-1) degrees of freedom when the null hypothesis is true, i.e.

$$Q^2 = \left(\boldsymbol{\psi}\,\hat{\boldsymbol{\eta}}\right)^{\mathsf{T}}\left(\boldsymbol{\psi}\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\eta}}}\boldsymbol{\psi}^{\mathsf{T}}\right)^{-1}\boldsymbol{\psi}\,\hat{\boldsymbol{\eta}} \xrightarrow[n\to\infty]{} \chi^2_{2(J-1)} \tag{3}$$

Alternative methods to the global hypothesis test based on the chi-squared distribution are the following:

1. Comparisons of the paired positive (negative) *LRs* of diagnostic tests to an error rate of α. This method consists of solving the 2J marginal hypothesis tests given by

$$\begin{aligned} H_0 : LR_k^+ = LR_l^+ \text{ vs } H_1 : LR_k^+ \neq LR_l^+ \\ H_0 : LR_k^- = LR_l^- \text{ vs } H_1 : LR_k^- \neq LR_l^- \end{aligned} \tag{4}$$

   when $k, l = 1, \ldots, J$ and $k \neq l$, each one of them to an error rate of α. Based on the asymptotic normality of the estimators of the *LRs*, the statistic for hypothesis test (4) is

$$z = \frac{\widehat{LR}_j - \widehat{LR}_k}{\sqrt{\widehat{\text{Var}}\left(\widehat{LR}_j\right) + \widehat{\text{Var}}\left(\widehat{LR}_k\right) - 2\widehat{\text{Cov}}\left(\widehat{LR}_j, \widehat{LR}_k\right)}} \xrightarrow[n\to\infty]{} N(0,1)$$

   where $\widehat{LR}$ is $\widehat{LR}^+$ or $\widehat{LR}^-$ and the variances-covariances are obtained from equation (1).

2. Another alternative method to the statistic (3) consists of solving the 2J marginal hypothesis tests (4) by applying a method of multiple comparisons, such as the Bonferroni method (1936), the Holm method (1979) or the Hochberg method (1983), which are very easy to apply and which are frequently used in the field of multiple comparisons. The Bonferroni method consists of solving each marginal hypothesis test (4) to an error rate of $\alpha/\{J(J-1)\}$ instead of to an error rate of α. In the Appendix there is a summary of the Holm method and the Hochberg method.

3. Simulation experiments

Monte Carlo simulation experiments were carried out to study the type I error and the power of the global hypothesis test based on the chi-squared distribution (3) and on the alternative methods proposed in the previous section, when simultaneously comparing the *LRs* of two and of three binary diagnostic tests respectively. The experiments consisted of the generation of 5000 random multinomial samples of size 100, 200, 300, 400, 500, 1000 and 2000. The samples were generated in such a way that for all of them it was possible to estimate the *LRs* and their variances-covariances. For all of the study we set $\alpha = 0.05$.

3.1. Two diagnostic tests

When simultaneously comparing the *LRs* of two binary diagnostic tests, as the sensitivity and specificity of each diagnostic test we took the values $\{0.70, 0.75, \ldots, 0.95\}$, which are values that frequently appear in clinical practice; as values for the disease prevalence we took $\{10\%, 20\%, 30\%, 40\%, 50\%\}$, and as the verification probabilities we took the values

$$(\lambda_{11} = 0.70,\ \lambda_{10} = \lambda_{01} = 0.40,\ \lambda_{00} = 0.10)$$

and

$$(\lambda_{11} = 0.95,\ \lambda_{10} = \lambda_{01} = 0.60,\ \lambda_{00} = 0.30),$$

that can be considered to be low and high verification probabilities respectively. The probabilities of the multinomial distributions were calculated applying Vacek's conditional dependence model (Vacek, 1985), i.e.

$$\begin{aligned}
\phi_{ij} &= \lambda_{ij} p \left\{ Se_1^i (1-Se_1)^{1-i} Se_2^j (1-Se_2)^{1-j} + \delta_{ij}\varepsilon_1 \right\}, \\
\varphi_{ij} &= \lambda_{ij} (1-p) \left\{ Sp_1^{1-i} (1-Sp_1)^i Sp_2^{1-j} (1-Sp_2)^j + \delta_{ij}\varepsilon_0 \right\}, \\
\gamma_{ij} &= (1-\lambda_{ij})\, p \left\{ Se_1^i (1-Se_1)^{1-i} Se_2^j (1-Se_2)^{1-j} + \delta_{ij}\varepsilon_1 \right\} \\
&\quad + (1-\lambda_{ij})(1-p) \left\{ Sp_1^{1-i} (1-Sp_1)^i Sp_2^{1-j} (1-Sp_2)^j + \delta_{ij}\varepsilon_0 \right\},
\end{aligned}$$

where $\delta_{ij} = 1$ when $i = j$ and $\delta_{ij} = -1$ when $i \neq j$, and ε_1 is the dependence factor (covariance) between the two diagnostic tests when $D = 1$ and ε_0 is the dependence factor (covariance) between the two diagnostic tests when $D = 0$. In general, in clinical practice the two diagnostic tests are usually conditionally dependent on the disease and it is verified that

$$0 < \varepsilon_1 < Se_1\,(1 - Se_2) \text{ when } Se_2 > Se_1$$
$$0 < \varepsilon_1 < Se_2\,(1 - Se_1) \text{ when } Se_1 > Se_2$$

and

$$0 < \varepsilon_0 < Sp_1\,(1 - Sp_2) \text{ when } Sp_2 > Sp_1$$
$$0 < \varepsilon_0 < Sp_2\,(1 - Sp_1) \text{ when } Sp_1 > Sp_2.$$

If the two diagnostic tests are conditionally independent on the disease then it is verified that $\varepsilon_1 = \varepsilon_0 = 0$.

In Table 1, we show the results obtained for the type I error when comparing the *LRs* of two diagnostic tests with sensitivities equal to 0.90 and specificities equal to 0.80, prevalence is equal to 10% and for intermediate and high dependence factors (ε_1 and ε_0). From the results, the following conclusions are obtained. The global hypothesis test based on the chi-squared distribution has a type I error which, in general terms, fluctuates around a nominal error of 5% especially when $n \geqslant 1000$, and the type I error is lower than the nominal error for samples of a smaller size. Therefore, the global test based on the chi-squared distribution show the classic performance of an asymptotic tests (the type I error fluctuates around the nominal error starting from a certain sample size). Moreover, the type I error increases when there is a rise in the disease prevalence but without overwhelming the nominal error of 5%, whilst the verification probabilities do not have an important effect upon the type I error (especially with large samples). Regarding the type I error of the method based on the paired comparison to an error rate of 5% (called Method 1 in the tables), its type I error clearly overwhelms the nominal error, above all when $n \geqslant 300 - 400$ depending on the prevalence and the verification probabilities, and therefore this method may lead to erroneous results. Regarding the methods based on paired comparisons and the application of the Bonferroni method (Method 2), the Holm method (Method 3) and the Hochberg method (Method 4), their respective type I errors are almost identical and show a very similar performance to the type I error of the global test based on the chi-squared distribution. Regarding power, in Table 2 we show the results obtained when the sensitivities are equal to 0.90 and 0.85 and the specificities are equal to 0.80 and 0.75 respectively, prevalence is equal to 10% and also for intermediate and high dependence factors. In general terms, with samples of 500 individuals, the power of the global test is very high (higher than 80%-90%), and the power is greater when the prevalence is greater and also when the verification probabilities are greater. Regarding the power of Method 2, this is greater than that of the global test because its type I error is also greater. As for the powers of Methods 2, 3 and 4, these are very similar to each other and these methods also have a power which is slightly lower than that of the global test, especially when the samples are not very large (in general terms, between 200 and 400 individuals).

Table 1: *Type I errors when comparing the LRs of two diagnostic tests.*

$Se_1 = Se_2 = 0.90$ $Sp_1 = Sp_2 = 0.80$ $p = 10\%$

$LR_1^+ = LR_2^+ = 4.5$ $LR_1^- = LR_2^- = 0.125$

$\lambda_{11} = 0.70$ $\lambda_{10} = 0.40$ $\lambda_{01} = 0.40$ $\lambda_{00} = 0.10$

	$\varepsilon_1 = 0.04$ $\varepsilon_0 = 0.07$					$\varepsilon_1 = 0.08$ $\varepsilon_0 = 0.14$				
n	**Global test**	**Method 1**	**Method 2**	**Method 3**	**Method 4**	**Global test**	**Method 1**	**Method 2**	**Method 3**	**Method 4**
100	0.007	0.010	0.005	0.005	0.005	0.000	0.000	0.000	0.000	0.000
200	0.006	0.008	0.005	0.005	0.005	0.001	0.004	0.002	0.002	0.003
300	0.005	0.010	0.000	0.000	0.000	0.008	0.010	0.008	0.008	0.009
400	0.010	0.020	0.007	0.007	0.008	0.018	0.022	0.015	0.015	0.016
500	0.011	0.025	0.009	0.009	0.010	0.013	0.020	0.012	0.012	0.013
1000	0.037	0.054	0.027	0.027	0.029	0.014	0.033	0.011	0.011	0.012
2000	0.044	0.086	0.040	0.040	0.041	0.025	0.052	0.026	0.026	0.027

$\lambda_{11} = 0.95$ $\lambda_{10} = 0.60$ $\lambda_{01} = 0.60$ $\lambda_{00} = 0.30$

	$\varepsilon_1 = 0.04$ $\varepsilon_0 = 0.07$					$\varepsilon_1 = 0.08$ $\varepsilon_0 = 0.14$				
n	**Global test**	**Method 1**	**Method 2**	**Method 3**	**Method 4**	**Global test**	**Method 1**	**Method 2**	**Method 3**	**Method 4**
100	0.004	0.011	0.003	0003	0.003	0.001	0.002	0.001	0.001	0.002
200	0.006	0.022	0.005	0.005	0.006	0.007	0.022	0.010	0.010	0.011
300	0.010	0.041	0.011	0.011	0.013	0.007	0.028	0.010	0.010	0.011
400	0.025	0.043	0.022	0.022	0.023	0.013	0.036	0.014	0.014	0.016
500	0.035	0.053	0.034	0.034	0.034	0.014	0.038	0.015	0.015	0.015
1000	0.043	0.080	0.040	0.040	0.042	0.022	0.056	0.022	0.022	0.024
2000	0.052	0.098	0.048	0.048	0.051	0.030	0.067	0.023	0.023	0.025

Table 2: *Powers when comparing the LRs of two diagnostic tests.*

$Se_1 = 0.90\ Se_2 = 0.85\ Sp_1 = 0.80\ Sp_2 = 0.75\ p = 10\%$

$LR_1^+ = 6\ LR_2^+ = 3.2\ LR_1^- = 0.12\ LR_2^- = 0.27$

$\lambda_{11} = 0.70\ \lambda_{10} = 0.40\ \lambda_{01} = 0.40\ \lambda_{00} = 0.10$

	$\varepsilon_1 = 0.03\ \varepsilon_0 = 0.05$					$\varepsilon_1 = 0.06\ \varepsilon_0 = 0.10$				
n	**Global test**	**Method 1**	**Method 2**	**Method 3**	**Method 4**	**Global test**	**Method 1**	**Method 2**	**Method 3**	**Method 4**
100	0.044	0.045	0.023	0.023	0.024	0.036	0.030	0.020	0.020	0.020
200	0.166	0.208	0.086	0.086	0.091	0.324	0.261	0.171	0.171	0.176
300	0.473	0.492	0.308	0.308	0.322	0.659	0.638	0.466	0.466	0.476
400	0.715	0.732	0.560	0.560	0.568	0.841	0.885	0.791	0.791	0.799
500	0.874	0.876	0.767	0.767	0.767	0.950	0.968	0.939	0.939	0.939
1000	1	0.995	0.992	0.992	0.992	1	1	1	1	1
2000	1	1	1	1	1	1	1	1	1	1

$\lambda_{11} = 0.95\ \lambda_{10} = 0.60\ \lambda_{01} = 0.60\ \lambda_{00} = 0.30$

	$\varepsilon_1 = 0.03\ \varepsilon_0 = 0.05$					$\varepsilon_1 = 0.06\ \varepsilon_0 = 0.10$				
n	**Global test**	**Method 1**	**Method 2**	**Method 3**	**Method 4**	**Global test**	**Method 1**	**Method 2**	**Method 3**	**Method 4**
100	0.069	0.092	0.039	0.039	0.047	0.131	0.135	0.052	0.052	0.069
200	0.340	0.505	0.315	0.315	0.324	0.647	0.748	0.618	0.618	0.622
300	0.738	0.809	0.678	0.678	0.680	0.936	0.946	0.901	0.901	0.901
400	0.907	0.923	0.856	0.856	0.856	0.922	0.990	0.980	0.980	0.980
500	0.974	0.972	0.941	0.941	0.941	0.998	0.998	0.998	0.998	0.998
1000	1	1	1	1	1	1	1	1	1	1
2000	1	1	1	1	1	1	1	1	1	1

3.2. Three diagnostic tests

When simultaneously comparing the *LRs* of three binary diagnostic tests, as the sensitivity and the specificity of each diagnostic test and the disease prevalence we took the same values as in the case of the diagnostic tests, and as verification probabilities we took the values

$$(\lambda_{111}=0.70,\lambda_{110}=0.40,\lambda_{101}=0.40,\lambda_{100}=0.25,\lambda_{011}=0.40,\lambda_{010}=0.25\,,\\ \lambda_{001}=0.25,\lambda_{000}=0.05)$$

and

$$(\lambda_{111}=1,\lambda_{110}=0.80,\lambda_{101}=0.80,\lambda_{100}=0.40,\lambda_{011}=0.80,\lambda_{010}=0.40\,,\\ \lambda_{001}=0.40,\lambda_{000}=0.20)$$

which can also be considered to be low and high verification scenarios. When comparing the *LRs* of three diagnostic tests, the probabilities of the multinomial distributions were calculating applying the Torrance-Rynard and Walter model (1997). In this case, the expressions of the probabilities are:

$$\phi_{i_1i_2i_3}=p\lambda_{i_1i_2i_3}\left\{\prod_{j=1}^{3}Se_j^{\,i_j}(1-Se_j)^{1-i_j}+\sum_{j,k,j<k}^{3}(-1)^{|i_j-i_k|}\delta_{jk}\right\},$$

$$\varphi_{i_1i_2i_3}=q\lambda_{i_1i_2i_3}\left\{\prod_{j=1}^{3}Sp_j{}^{1-i_j}(1-Sp_j)^{i_j}+\sum_{j,k,j<k}^{3}(-1)^{|i_j-i_k|}\varepsilon_{jk}\right\},$$

$$\gamma_{i_1i_2i_3}=p\left(1-\lambda_{i_1i_2i_3}\right)\left\{\prod_{j=1}^{3}Se_j{}^{i_j}(1-Se_j)^{1-i_j}+\sum_{j,k,j<k}^{3}(-1)^{|i_j-i_k|}\delta_{jk}\right\}$$

$$+(1-p)\left(1-\lambda_{i_1i_2i_3}\right)\left\{\prod_{j=1}^{3}Sp_j{}^{1-i_j}(1-Sp_j)^{i_j}+\sum_{j,k,j<k}^{3}(-1)^{|i_j-i_k|}\varepsilon_{jk}\right\},$$

with $i_j=0,1$, $i_k=0,1$ and $j,k=1,2,3$, where δ_{jk} is the dependence factor (covariance) between the *j*th and the *k*th diagnostic test when $D=1$ and ε_{jk} is the dependence factor (covariance) between the *j*th and the *k*th diagnostic test when $D=0$. The dependence factors δ_{jk} and ε_{jk} verifies restrictions that depend on the values of the sensitivities and the specificities of the three diagnostic tests. In order to simplify things, in the simulation experiments it was considered that $\delta_{ij}=\delta$ and $\varepsilon_{ij}=\varepsilon$, and therefore the dependence factors verify the following restrictions:

$$\delta\leqslant(1-Se_1)(1-Se_2)Se_3,\ \delta\leqslant(1-Se_1)Se_2(1-Se_3),\ \delta\leqslant Se_1(1-Se_2)(1-Se_3)$$
$$\varepsilon\leqslant(1-Sp_1)(1-Sp_2)Sp_3,\ \varepsilon\leqslant(1-Sp_1)Sp_2(1-Sp_3),\ \varepsilon\leqslant Sp_1(1-Sp_2)(1-Sp_3).$$

Table 3: *Type I errors when comparing the LRs of three diagnostic tests.*

$Se_1 = Se_2 = Se_3 = 0.90\ Sp_1 = Sp_2 = Sp_3 = 0.80\ p = 10\%$

$LR_1^+ = LR_2^+ = LR_3^+ = 4.5\ LR_1^- = LR_2^- = LR_3^- = 0.125$

$\lambda_{111} = 0.70\ \lambda_{110} = 0.40\ \lambda_{101} = 0.40\ \lambda_{100} = 0.25\ \lambda_{011} = 0.40\ \lambda_{010} = 0.25\ \lambda_{001} = 0.25\ \lambda_{000} = 0.05$

	$\delta = 0.004\ \varepsilon = 0.015$					$\delta = 0.008\ \varepsilon = 0.03$				
n	**Global test**	**Method 1**	**Method 2**	**Method 3**	**Method 4**	**Global test**	**Method 1**	**Method 2**	**Method 3**	**Method 4**
100	0.001	0.003	0.000	0.000	0.001	0.000	0.001	0.000	0.000	0.000
200	0.005	0.022	0.002	0.002	0.014	0.004	0.028	0.001	0.001	0.001
300	0.020	0.078	0.012	0.012	0.014	0.004	0.028	0.001	0.001	0.001
400	0.036	0.123	0.023	0.023	0.024	0.009	0.053	0.007	0.007	0.007
500	0.043	0.140	0.028	0.029	0.032	0.021	0.078	0.011	0.011	0.012
1000	0.054	0.199	0.034	0.034	0.040	0.045	0.179	0.035	0.035	0.039
2000	0.056	0.216	0.052	0.052	0.057	0.058	0.216	0.048	0.049	0.051

$\lambda_{111} = 1\ \lambda_{110} = 0.80\ \lambda_{101} = 0.80\ \lambda_{100} = 0.40\ \lambda_{011} = 0.80\ \lambda_{010} = 0.40\ \lambda_{001} = 0.40\ \lambda_{000} = 0.20$

	$\delta = 0.004\ \varepsilon = 0.015$					$\delta = 0.008\ \varepsilon = 0.03$				
n	**Global test**	**Method 1**	**Method 2**	**Method 3**	**Method 4**	**Global test**	**Method 1**	**Method 2**	**Method 3**	**Method 4**
100	0.010	0.013	0.001	0.001	0.001	0.001	0.004	0.001	0.001	0.001
200	0.019	0.077	0.005	0.005	0.007	0.004	0.038	0.003	0.003	0.004
300	0.028	0.125	0.013	0.013	0.016	0.012	0.087	0.005	0.005	0.006
400	0.032	0.148	0.018	0.018	0.020	0.023	0.125	0.014	0.014	0.016
500	0.039	0.161	0.021	0.021	0.024	0.035	0.157	0.019	0.019	0.021
1000	0.051	0.204	0.036	0.036	0.039	0.051	0.189	0.039	0.040	0.042
2000	0.053	0.221	0.041	0.041	0.046	0.050	0.204	0.037	0.037	0.042

Table 4: *Powers when comparing the LRs of three diagnostic tests.*

$Se_1 = 0.90\ Se_2 = 0.85\ Se_3 = 0.80\ Sp_1 = 0.85\ Sp_2 = 0.75\ Sp_3 = 0.70\ p = 10\%$

$LR_1^+ = 6\ LR_2^+ = 3.4\ LR_3^+ = 2.67\ LR_1^- = 0.12\ LR_2^- = 0.2\ LR_3^- = 0.29$

$\lambda_{111} = 0.70\ \lambda_{110} = 0.40\ \lambda_{101} = 0.40\ \lambda_{100} = 0.25\ \lambda_{011} = 0.40\ \lambda_{010} = 0.25\ \lambda_{001} = 0.25\ \lambda_{000} = 0.05$

	$\delta = 0.005\ \varepsilon = 0.01$					$\delta = 0.01\ \varepsilon = 0.02$				
n	**Global test**	**Method 1**	**Method 2**	**Method 3**	**Method 4**	**Global test**	**Method 1**	**Method 2**	**Method 3**	**Method 4**
100	0.007	0.006	0.001	0.001	0.001	0.005	0.007	0.001	0.001	0.001
200	0.194	0.296	0.080	0.080	0.088	0.123	0.200	0.060	0.060	0.067
300	0.617	0.739	0.411	0.412	0.425	0.535	0.610	0.395	0.396	0.408
400	0.868	0.907	0.667	0.668	0.682	0.831	0.856	0.708	0.709	0.719
500	0.949	0.967	0.822	0.824	0.836	0.909	0.945	0.833	0.833	0.845
1000	0.998	1	0.998	0.998	0.998	0.996	1	0.998	0.998	0.999
2000	1	1	1	1	1	1	1	1	1	1

$\lambda_{111} = 1\ \lambda_{110} = 0.80\ \lambda_{101} = 0.80\ \lambda_{100} = 0.40\ \lambda_{011} = 0.80\ \lambda_{010} = 0.40\ \lambda_{001} = 0.40\ \lambda_{000} = 0.20$

	$\delta = 0.005\ \varepsilon = 0.01$					$\delta = 0.01\ \varepsilon = 0.02$				
n	**Global test**	**Method 1**	**Method 2**	**Method 3**	**Method 4**	**Global test**	**Method 1**	**Method 2**	**Method 3**	**Method 4**
100	0.077	0.181	0.017	0.017	0.021	0.062	0.168	0.010	0.010	0.013
200	0.463	0.822	0.364	0.365	0.395	0.474	0.818	0.427	0.430	0.462
300	0.849	0.969	0.790	0.790	0.801	0.860	0.975	0.851	0.852	0.863
400	0.963	0.992	0.943	0.943	0.946	0.975	0.997	0.972	0.973	0.975
500	0.995	0.999	0.983	0.983	0.984	0.999	1	0.992	0.992	0.992
1000	1	1	1	1	1	1	1	1	1	1
2000	1	1	1	1	1	1	1	1	1	1

In clinical practice, factors δ_{jk} and/or ε_{jk} are greater than zero, and therefore the diagnostic tests are conditionally dependent on the disease status. When $\delta_{jk} = \varepsilon_{jk} = 0$ the three diagnostic tests are conditionally independent on the disease status.

In Table 3, we show the results obtained for the type I error when the three sensitivities are equal to 0.90 and the three specificities are equal to 0.80, prevalence is equal to 10% and for intermediate and high dependence factors (δ and ε). From the results it holds that, in general terms, the type I error of the global hypothesis test performs in a similar way to that obtained when comparing two diagnostic tests (the type I error fluctuates around the nominal error starting from a determined sample size). Regarding the other methods, Method 1 has a type I error that clearly overwhelms the nominal error, and Methods 2, 3 and 4 have a type I error that is slightly lower than that of the global test.

In terms of power, in Table 4 we show the results obtained for sensitivities equal to 0.90, 0.85 and 0.80, specificities equal to 0.85, 0.75 and 0.70 respectively, and prevalence is equal to 10%. In general terms, the power of the global test increases with an increase in the prevalence and/or the verification probabilities, and the power is greater than 80%-90% with samples of 500. Furthermore, Method 1 has a greater power than the global test because (as in the case of the two diagnostic tests) its type I error is greater. Methods 2, 3 and 4 have a power which is slightly lower than the global test, especially for samples of between 100 and 400 individuals.

3.3. Conclusions

From the results of the simulation experiments carried out to simultaneously compare the *LRs* of two and three diagnostic tests respectively, it holds that in general terms the best method to solve this problem of inference is the global test based on the chi-squared distribution, since its type I error performs better around the nominal error than the type I error of each one the other methods. From these results, the following method is proposed to compare the likelihood ratios of J binary diagnostic tests: 1) Solving the global hypothesis test based on the chi-squared distribution to an error rate of α; 2) If the global hypothesis is significant to an error rate of α, the investigation of the causes of the significance must be carried out comparing the positive (negative) likelihood ratios of each pair of diagnostic tests applying a multiple comparison method (Bonferroni, Holm or Hochberg) to an error α. Step 2 must be carried out applying a multiple comparison method and not each marginal test to an error rate of α, since the latter has a type I error that clearly overwhelms the nominal error.

4. Application

The results obtained in previous Sections were applied to the diagnosis of coronary stenosis, a disease that consists of the obstruction of the coronary artery and its diagnosis can be made through a dobutamine echocardiography, a stress echocardiography or through a *CT* scan, and as the gold standard a coronary angiography is used. As the coronary angiography can cause different reactions in individuals (thrombosis, heart attack, infections, etc.), not all of the individuals are verified with the coronary angiography. In Table 5, we show the results obtained when applying the three diagnostic tests and the gold standard (T_1: dobutamine ecocardiography; T_2: stress echocardiography; T_3: *CT* scan) to a sample of 2455 spanish males over 45 and when applying the coronary angiography (D) only to a subset of these individuals. The data come from a study carried out at the University Hospital in Granada. This study was carried out in two phases: in the first phase, the three diagnostic tests were applied to all of the individuals; and in the second phase, the coronary angiography was applied only to a subset of these individuals depending only on the results of the three diagnostic tests. Therefore, in this example it can be assumed that the missing data mechanism is *MAR* and the model is ignorable, and therefore the results of the previous sections can be applied. The values of the estimators of the *LRs* are $\widehat{LR}_1^+ = 5.31$, $\widehat{LR}_2^+ = 3.04$, $\widehat{LR}_3^+ = 7.61$, $\widehat{LR}_1^- = 0.13$, $\widehat{LR}_2^- = 0.33$ and $\widehat{LR}_3^- = 0.09$. Applying equation (3) it holds that $Q^2 = 126.20$ (p-value $= 0$) and therefore we reject the joint equality of the *LRs*. In order to investigate the causes of the significance, the step is to solve the marginal hypothesis tests. In Table 6, we show the results obtained for each one of the six hypothesis tests that compare the *LRs*. Then a method of multiple comparisons (Bonferroni, Holm or Hochberg) is applied and it is found that (with the three methods) the three positive likelihood ratios are different, and the biggest one is that of the *CT* scan, followed by that of the dobutamine echocardiography and finally that of the stress echocardiography. Regarding the negative likelihood ratios, no significant differences were found between that of the dobutamine echocardiography and that of the *CT* scan; whilst the negative likelihood ratio of the stress echocardiography is significantly larger than that of the dobutamine echocardiography and that of the *CT* scan.

5. Discussion

Likelihood ratios are parameters that allow us to assess and compare the performance of binary tests, and technically they are equivalent to a relative risk. In the presence of partial disease verification, the disease status of a subset of individuals in the sample is unknown, and therefore the estimation and comparison of the likelihood ratios of two or more diagnostic tests cannot be made using the existing models (Leisenring and Pepe, 1998; Roldán Nofuentes and Luna del Castillo, 2007), since the results are affected by verification bias. In this article, a global hypothesis test is proposed to simultaneously

Table 5: *Data from the study of coronary stenosis.*

	$T_1=1$				$T_1=0$				
	$T_2=1$		$T_2=0$		$T_2=1$		$T_2=0$		
	$T_3=1$	$T_3=0$	$T_3=1$	$T_3=0$	$T_3=1$	$T_3=0$	$T_3=1$	$T_3=0$	Total
$V=1$									
$D=1$	457	30	84	5	34	0	7	1	618
$D=0$	41	23	5	61	16	86	32	95	359
$V=0$	92	31	85	120	42	195	88	825	1478
Total	590	84	174	186	92	281	127	921	2455

Table 6: *Results of the marginal hypothesis tests.*

Hypothesis test	z	Two sided p-value
$H_0: LR_1^+ = LR_2^+$ vs $H_1: LR_1^+ \neq LR_2^+$	6.24	4.47×10^{-13}
$H_0: LR_1^+ = LR_3^+$ vs $H_1: LR_1^+ \neq LR_3^+$	3.30	0.001
$H_0: LR_2^+ = LR_3^+$ vs $H_1: LR_2^+ \neq LR_3^+$	7.29	3.06×10^{-13}
$H_0: LR_1^- = LR_2^-$ vs $H_1: LR_1^- \neq LR_2^-$	7.53	5.15×10^{-14}
$H_0: LR_1^- = LR_3^-$ vs $H_1: LR_1^- \neq LR_3^-$	1.77	0.077
$H_0: LR_2^- = LR_3^-$ vs $H_1: LR_2^- \neq LR_3^-$	9.19	0

compare the likelihood ratios of two or more diagnostic tests assuming that the missing data mechanism is ignorable. The assumption of ignorability (Schafer, 1997), which is widely used in this field, means that the selection of an individual to verify the disease status depends only on the results of the diagnostic tests and not on the disease status. This assumption cannot be made from the data observed, but rather depends on how the experiment is conducted. Thus, for example, in two phase studies, if in the second phase the selection of the individuals is made depending on the results of the diagnostic tests, then it can be assumed that the missing data mechanism is ignorable. If the verification process depends on the disease status, the missing data mechanism is not ignorable and the model proposed in this article cannot be applied.

Simulation experiments were carried out to study the type I error and the power of the global test and of other alternative methods, from which the following method was proposed to compare the likelihood ratios of two or more diagnostic tests in the presence of ignorable missing data: 1) Apply the global hypothesis test based on the chi-squared distribution to an error rate of α (equation (3)); 2) If the global hypothesis test is significant to an error rate of α, investigating the causes of the significance solving the marginal hypothesis tests (expression (4)) along with the a multiple comparison method (Bonferroni, Holm or Hochberg). This procedure is similar to the one used in a analysis of variance. Firstly, the global test is solved and then a multiple comparison method is applied. The simulation experiments have also shown that the positive

and negative likelihood ratios cannot be compared independently (Method 1 of the simulation experiments), since the type I error clearly overwhelms the nominal error.

The problem posed in this article can also be solved using the natural log-likelihood ratios. Simulation experiments (similar to those in Section 3 and from the same samples) were carried out using this transformation and it was found that there is no important difference between the results obtained, in terms of type I error and power, and those obtained in Section 3. Therefore, it is recommendable to make the comparison without using this transformation.

Acknowledgements

This research was supported by the Spanish Ministry of Science, Grant Number MTM 2012-35591. We thank the two referees, the Associate Editors (Montserrat Guillén Estany and David Conesa) of SORT for their helpful comments that improved the quality of the paper.

Appendix

Let us suppose that we wish to check K hypothesis tests, H_{0k} vs H_{1k}, with $k = 1,\ldots,K$, and let p_k be the p-value obtained by solving each hypothesis test. Let $p_{[1]} \leqslant p_{[2]} \leqslant \cdots \leqslant p_{[K]}$ be the p-values in order from the lowest to the highest, so that $p_{[k]}$ is the p-values corresponding to the hypothesis test $H_{0[k]}$ vs $H_{1[k]}$.

The Holm method (1979) consists of the following steps:

Step 1. If $p_{[1]} \leqslant \alpha/K$ then reject hypothesis $H_{0[1]}$ and go to the next step; otherwise finish.

Step 2. If $p_{[2]} \leqslant \alpha/(K-1)$ then reject hypothesis $H_{0[2]}$ and go to the next step; otherwise finish...

Step K. If $p_{[K]} \leqslant \alpha$ then reject hypothesis $H_{0[K]}$ and finish.

The Hochberg method (1988) consists of the following steps:

Step 1. If $p_{[K]} \leqslant \alpha$ then reject $H_{0[k]}$ with $k = 1,\ldots,K$ and finish; otherwise go to the next step.

Step 2. If $p_{[K-1]} \leqslant \alpha/2$ then reject $H_{0[k]}$ with $k = 1,\ldots,K-1$ and finish; otherwise go to the next step...

Step K. If $p_{[1]} \leqslant \alpha/K$ then reject hypothesis $H_{0[1]}$ and finish.

References

Agresti, A. (2002). *Categorical Data Analysis*. New York: Wiley.

Begg, C. B. and Greenes, R. A. (1983). Assessment of diagnostic tests when disease verification is subject to selection bias. *Biometrics*, 39, 207–215.

Bonferroni, C. E. (1936). Teoria statistica delle classi e calcolo delle probabilità. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, 8, 3–62.

Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biométrica*, 75, 800–802.

Holm, S. (1979). A simple sequential rejective multiple testing procedure. Scandinavian. *Journal of Statistics*, 6, 65–70.

Leisenring, W. and Pepe, M. S. (1998). Regression modelling of diagnostic likelihood ratios for the evaluation of medical diagnostic tests. *Biometrics*, 54, 444–452.

Luts, J., Roldán Nofuentes, J. A., Luna del Castillo, J. D. and Van Huffel, S. (2011). Asymptotic hypothesis test to compare likelihood ratios of multiple diagnostic tests in unpaired designs. *Journal of Statistical Planning and Inference*, 141, 3578–3594.

Roldán Nofuentes, J. A. and Luna del Castillo, J. D. (2005). Comparing the likelihood ratios of two binary diagnostic tests in the presence of partial verification. *Biometrical Journal*, 47, 442–457.

Roldán Nofuentes, J. A. and Luna del Castillo, J. D. (2007). Comparison of the likelihood ratios of two binary diagnostic tests in paired designs. *Statistics in Medicine*, 26, 4179–4201.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 4, 73–89.

Schafer, J. L. (1997). Analysis of Incomplete Multivariate Data. USA: Chapman and Hall/CRC.

Torrance-Rynard, V. L. and Walter, S. D. (1997). Effects of dependent errors in the assessment of diagnostic test performance. *Statistics in Medicine*, 16, 2157–2175.

Vacek, P. (1985). The effect of conditional dependence on the evaluation of diagnostic tests. *Biometrics*, 41, 959–968.

Zhou, X. H. (1993). Maximum likelihood estimators of sensitivity and specificity corrected for verification bias. *Communication in Statistics-Theory and Methods*, 22, 3177–3198.